

Predicting the Brand Popularity from the Brand Metadata

Bhargavi K¹, Sathish Babu B², S. S. Iyengar³

¹Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

²Department of Computer Science and Engineering, R V College of Engineering, Bengaluru, Karnataka, India

³Florida International University, Miami, Florida, USA

Article Info

Article history:

Received May 21, 2018

Revised Aug 20, 2018

Accepted Aug 28, 2018

Keyword:

Brand metadata

Brand popularity

MapReduce

Social networks

Thoughtful comment

ABSTRACT

Social networks have become one of the primary sources of big data, where a variety of posts related to brands are liked, shared, and commented, which are collectively called as brand metadata. Due to the increased boom in E/M-commerce, buyers often refer the brand metadata as a valuable source of information to make their purchasing decision. From the literature study, we found that there are not many works on predicting the popularity of the brand based on the combination of brand metadata and comment's thoughtfulness analysis. This paper proposes a novel framework to classify the comment's as thoughtful favored or disfavored comment's, and later combines them with the brand metadata to forecast the popularity of the brand in near future. The performance of the proposed framework is compared with some of the recent works w.r.t. thoughtful comment's identification accuracy, execution time, prediction accuracy and prediction time, the results obtained are found to be very encouraging.

Copyright © 2018 Institute of Advanced Engineering and Science.

All rights reserved.

Corresponding Author:

Bhargavi K,

Faculty of Computer Science and Engineering,

Siddaganga Institute of Technology,

Tumkur, India.

Phone: +91-9886280931

Email: bhargavi.tumkur@gmail.com

1. INTRODUCTION

Social networks have connected billions of people all over the world, who generate big data in the form of text, image, and audio/video. This data would serve as a valuable source of information for many big data researchers [1]-[3]. Due to increased penetration of socialization in daily lives, the social networking has turned out to be a prominent platform for brand advertisements. This advertisement could be personalized based on the customer profile, demography specific interests, customer feedback, and other parameters. It has been analyzed that advertisers will spend over 50 billion dollars on social media advertising by 2020 [4]-[6]. Since subscribers have the freedom to express their opinions on social network sites, the platform can be misused to post meaningless or not-thoughtful comments over the brands [7]-[10].

In this work, a framework is proposed to predict the brand popularity based on brand metadata and comment's analysis. This framework identifies thoughtful comments from the brand comment corpus and uses the comments to evaluate the current popularity of the brand. Then perform predictive analytics on the number of likes, the number of shares, and the number of identified thoughtful comments to predict the brand popularity status in the near future. Overall the brand popularity prediction aims to answer the question, i.e., "What popularity level the Brand B will be at future time T?". The proposed thoughtful comments identifier preprocesses the comments using Apache OpenNLP parser, and the `opennlp.grok.ml.dectree` class is used to identify the thoughtful comments [11]. Other metadata fields like number of likes, number of shares are combined with the number of identified thoughtful comments to forecast the popularity of the brand in near future.

The process of identifying thoughtful comments from the social media is discussed in [12], here the problems related to the thoughtful comments identification is discussed first then the textual features of the comments along with the discourse relationship among the words in the comments are used to predict the thoughtful comments using logistic regression model. The accuracy of the prediction falls below average as the model fails to identify the non-linear textual relationship in the large corpus of comments.

Online contents based popularity prediction model was constructed in [13]; the Cox proportional hazard regression model was used for prediction purpose by considering publicly available metrics of online content like thread lifetime, the number of comments, and the number of views. However, the popularity prediction was rendered by considering only the publicly available metrics during the first hour of the online content publication, which limits the scalability of the approach.

Attention prediction on brand pages in social media was discussed in [14], both content and network features of the generated comment's were used to determine the brand popularity. The attention gained by the brand post was determined using regression and classification methodology. However, detailed discussion on the aggregated analysis of user-specific interest and its influence on the popularity of brand pages were not provided.

The popularity of the news item in social media was discussed in [15]. The social media considered was Twitter; the popularity of the news item was determined by extracting content features from the news articles. The multidimensional feature space model for every news item in articles was constructed, which served as a prominent indicator of brand popularity. Still, the influence of individual propagators on brand popularity was not clearly discussed.

The [16], focuses mainly on the socio-political area, which highlights the issues related to mining the data from the social web, extraction of opinions related to the topic, and identification of thoughtful opinions from the comment's available in the socio-political websites. Here the Kullback-Leibler (KL) divergence algorithm was used to determine relevance words in the comments, but its performance was found to be lower when user abbreviated words in the comments as the algorithm did not use any topic modeling to check relevance between the comments.

The popularity of the news in social media is predicted in [17] on the basis of the number of likes, shares, and the comment's the article gets prior to the publication. A gradient boosting machine is developed to predict the popularity of the article in near future by using the metadata of the article. Here, for prediction purpose, the original metadata available after article publication is not taken into consideration and the comment's are not verified for thoughtfulness so the prediction rate falls on a decent scale.

A proactive system is proposed to forecast the popularity of the online news in [18], the rolling window evaluation followed by hill climbing local search is explored on a large collection of news dataset then random forest algorithm was run to classify the news content as popular or not popular. But the news contents collected for popularity prediction are static in nature and issues related to natural language processing of the news text are not taken into consideration this limits the practical applicability of the system.

The [19], describes several statistical self-learning frameworks useful for content popularity prediction. Two methods are used to assess the content popularity one is regression based and other is classification based. It has been said that the error rate in content popularity prediction is lower in the generalized additive mode regression model and random forecast classification model but both of the models suffers from overfitting problem due to the presence of noise in the content samples.

The popularity of the news articles is forecasted in [20], based on the article metadata, content, sentiment, readability, and named entity features. Here the problem of popularity prediction is considered as a regression problem and predicts the number of views of the article in the future. The only social media feature considered for popularity prediction was twitter, the social media-based features like the number of retweets, and the number of followers is used as popularity measurement metrics. The popularity prediction model suffers from scalability issue as only social media considered for analysis is twitter and all tweets are considered for analysis without investigating its thoughtfulness.

The factors affecting the popularity of the brand posts are surveyed in [21]. The factors identified are number of likes, number of comment's, and number of shares and the relationship between fans and brand posts are measured in terms of number of new likes and number of unlikes over the posts. The analysis shows that greeting posts received more number of likes, photos with messages received more number of shares, and weekend posts received more number of comments.

The influence of likes, shares, and comment's towards spreading of Facebook messages is described in [22]. The study reveals that if a person has liked a message then the chances of that person commenting or sharing the message will be high as there exists a high correlation between liking, sharing or commenting activities of the end users.

In literature, separate works exist with respect to thoughtful comments identification and brand popularity prediction purpose. The existing works regarding thoughtful comments identification fail to identify non-linear textual analysis in the comments and the existing works regarding popularity prediction suffers from overfitting problem. Hence the proposed work first identifies thoughtful comments and then uses it as one of the attributes along with the other metadata attributes of brand for prediction purpose which efficiently deals with both non-linear textual analysis and overfitting problem.

2. RESEARCH METHOD

Consider a Comment's Log $CL = \langle BM_1, BM_2, BM_3, \dots, BM_n \rangle$ consisting of comment's on several Brand Models (BMs), let $BM = \langle BP_1, BP_2, \dots, BP_n \rangle$ is a stereotype of brand containing several Brand Posts (BPs), and let $BP = \langle BP_{id}, BPC, N_l, N_s \rangle$ is a form of promotion comprising of Brand Post identifier (BP_{id}), set of Brand Post Comment's (BPC), Number-of-likes (N_l), and Number-of-shares (N_s).

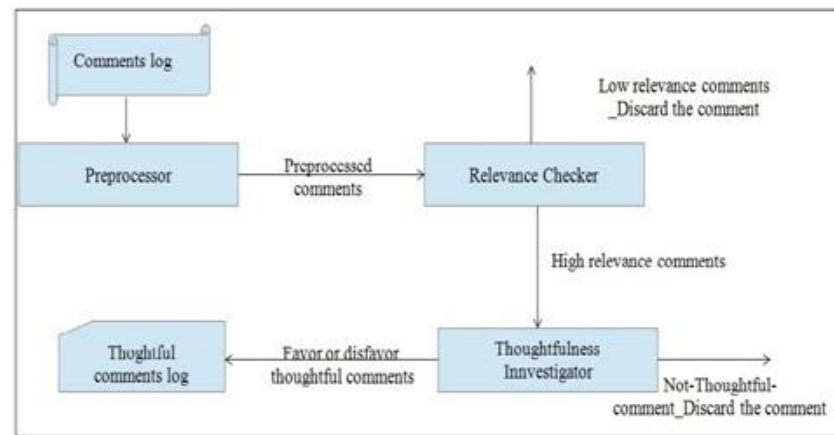


Figure 1. Thoughtful comments identifier framework

The Figure 1 depicts the framework used for thoughtful comments identification. The preprocessor cleans the BPC of every BPs in CL by removing stop words, noisy text, punctuation, and so on to yield preprocessed comments as output, relevance checker analyzes comments for their relevance level, and thoughtfulness investigator inspects the thoughtfulness of the comment. A Thoughtful Comment $TC = \langle Q, R, E, C \rangle$ comprises of thoughtful Questioning (Q), Reasoning (R), Experience (E), and Comparison (C) words, the words may be in favor or disfavor of the brand. The Thoughtful Favor Comment $TFC = \langle E \rightarrow +R / E \rightarrow +Q \rangle$ i.e., experience words followed by positive reasoning words or comparison words followed by positive questioning words, and Thoughtful Dis-Favor Comment $TDFC = \langle E \rightarrow -R / E \rightarrow -Q \rangle$ i.e., experience words followed by negative reasoning words or comparison words followed by negative questioning words are identified based on the sequence of appearance of thoughtful words in the comment. The sequence being considered is inspired by the thought process of the human beings while framing meaningful sentences. MapReduce model [23] is used to preprocess and classify the comment's as Uni-gram (one word in the comment), Bi-gram (two words in the comment), or N-gram (N words in the comment). The pictorial representation of classification along with an example is shown in Figure 2 and Figure 3.

As a sample case, we have considered over two hundred thousand comment posts on Hyundai cars for analysis. The posts include various models of Hyundai like Nguyen Van Sang, Hyundai i10, Hyundai i20, Azera, Genesis G90, Veloster Turbo, Sonata Hybrid, SantaFe, and so on. A scatter plot depicting the distribution of different type of comment's i.e., unigram, bigram, or multigram comments generated over various models of Hyundai with respect to time is shown in Figure 4(a). The preprocessed output of those comments is depicted in Figure 4(b), which shows that most of the comments were multigram by nature.

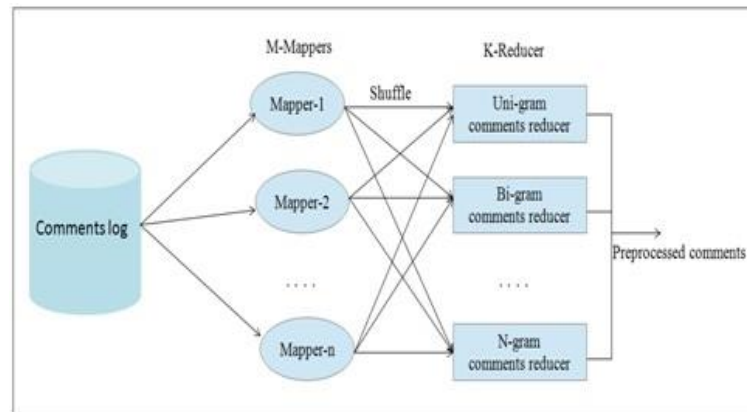


Figure 2. A sample MapReduce framework for comment preprocessing in preprocessor

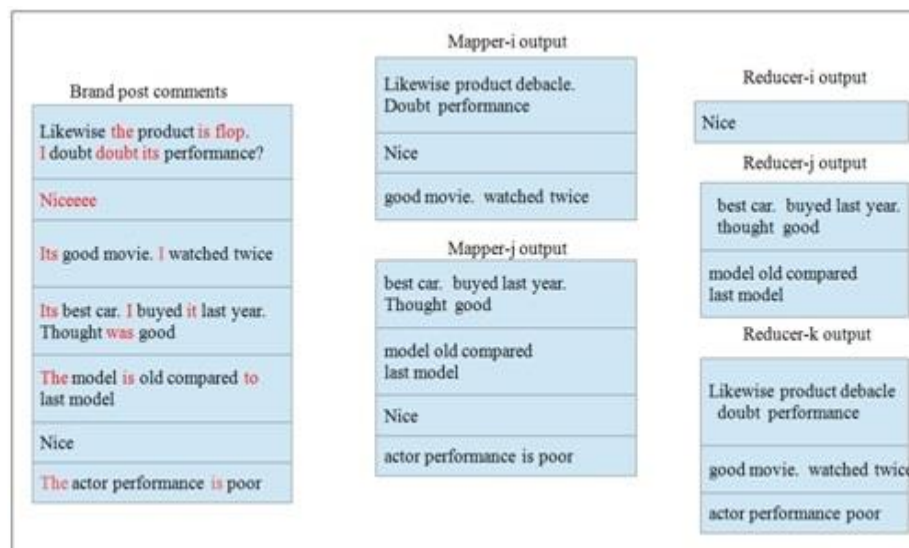


Figure 3. Comment preprocessing example

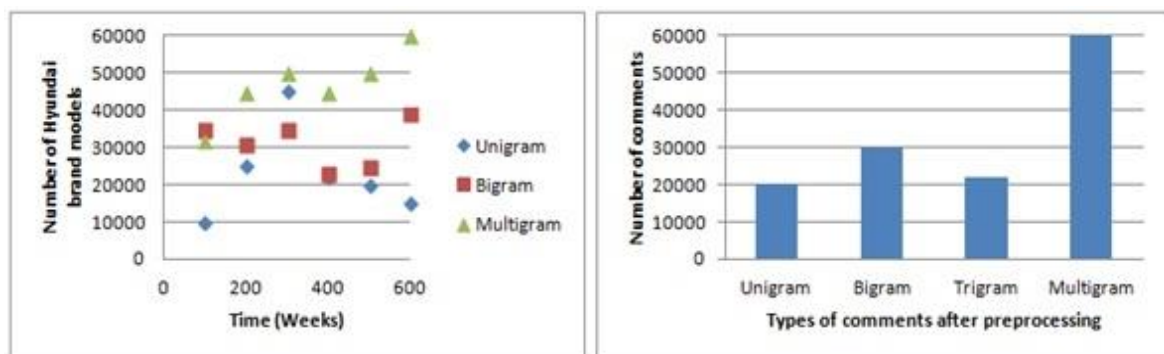


Figure 4. (a) Time versus number of hyundai brand models (b) Types of comments after preprocessing versus number of comments

In relevance checker, \forall BPid, and Preprocessed Comments PC_i in Preprocessed Comments set PC ; the lexical density is computed based on the number of lexical words in PC_i to produce High Lexical Density

Comments set *HLDC* [24]-[29]. The *HLDC* are further evaluated for relevance using the Latent Dirichlet allocation topic modeling technique. During topic modeling process, the Relevance Comments set *RC* is initialized with a list of relevance topic words from the comments. The probability of comments belonging to *RC* i.e., $P(\frac{RC}{HLDCi})$ is determined, and then every *HLDCi* exhibiting higher probability towards *HLDC* is aggregated to output $RC ::= RC \cup HLDCi$. An example for identifying relevant comments from *PC* is shown in Figure 5, and frequently occurring topic relevant words along with the topic irrelevant words with respect to Hyundai cars comments is shown in Figure 6 [11].

The *RC* is further classified into favored or disfavored by thoughtfulness investigator using decision tree prediction model. The decision tree logic works in two phases i.e., training and testing. During training phase; $\forall BPid$, and $RCi \in RC$, if the $Q \cup R \cup C \cup E$ is *NULL*, the comment is considered as not-thoughtful comment which is discarded, if the RCi exhibits positive sequence of words i.e., $E \rightarrow +RV C \rightarrow +Q$ then it is considered as *TFC* else if it exhibits the negative sequence of words i.e., $E \rightarrow -RV C \rightarrow -Q$ then it is considered as *TDFC*. During testing phase, $\forall BPid$ the *TFC* set $TFCS ::= TFCS \cup TFC$ and *TDFC* set $TDFCS ::= TDFCS \cup TDFC$ are enumerated. An example for identifying thoughtful comments from *HRC* is shown in Figure 7. The most frequently occurring positive and negative thoughtful words on Hyundai social media pages are given in Table 1.

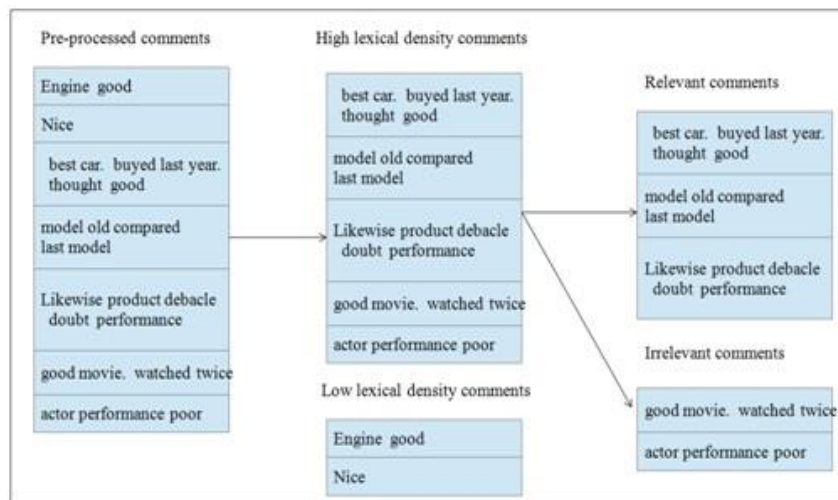


Figure 5 Example for identifying relevant comment

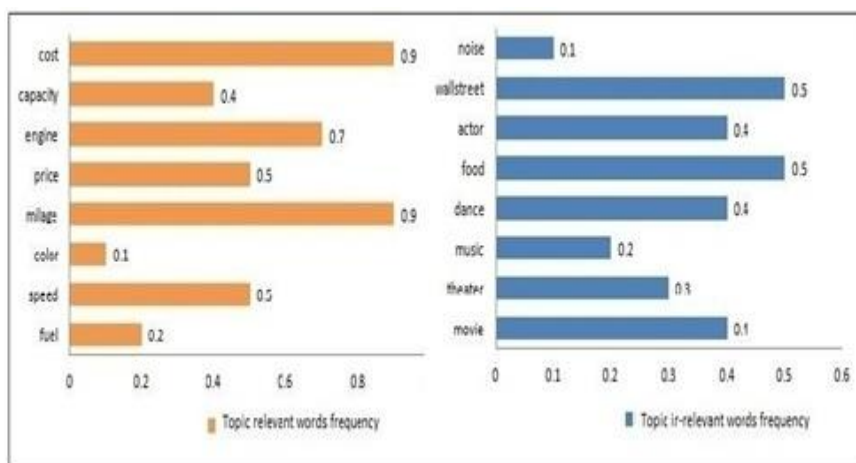


Figure 6 Topic relevant and irrelevant words

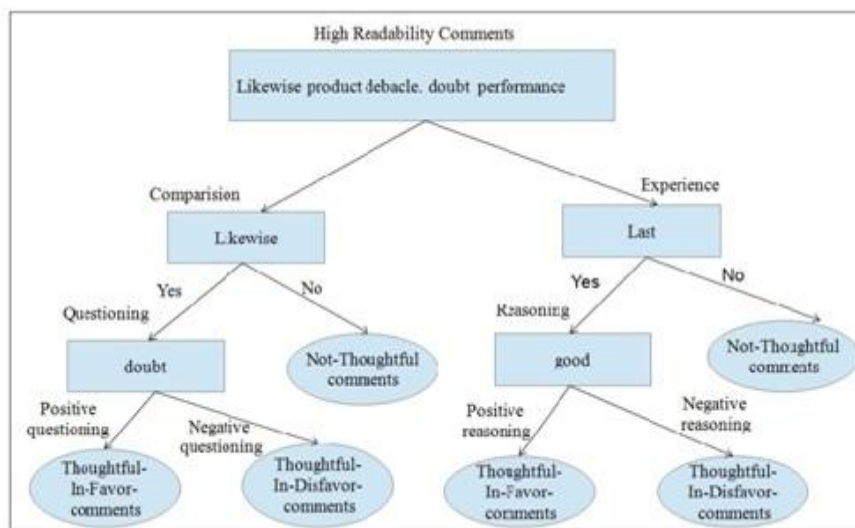


Figure 7. Thoughtful comments identification example

The thoughtfulness of the comment is highly influenced by the correlation between the number of words in the comment. It is measured on a scale of 0 to 1, is dependent on the number of words in the comment, as shown in the scatter plot in Figure 8(a). The number of thoughtful favor or disfavor comments on Hyundai cars during the interval 1/1/2016 to 1/3/2017 (120 weeks) is given in Figure 8(b). Two third of the thoughtful comments posted were disfavored as most of them expressed the negative opinion.

Table 1. Top 10 Extracted Words

Categories of words	Extracted words
Experience	trial, observation, sophisticate, practice, familiar, deal, spark, produce, check, tried
Comparison	similar, alike, correlate, analyze, example, estimation, measure, collate, together, like
Questioning	which, how, when, haven't, didn't, don't, does it, wont, would, shall
Reasoning	clear, obvious, thought, opinion, think, limit, argue, also, think, conclusion

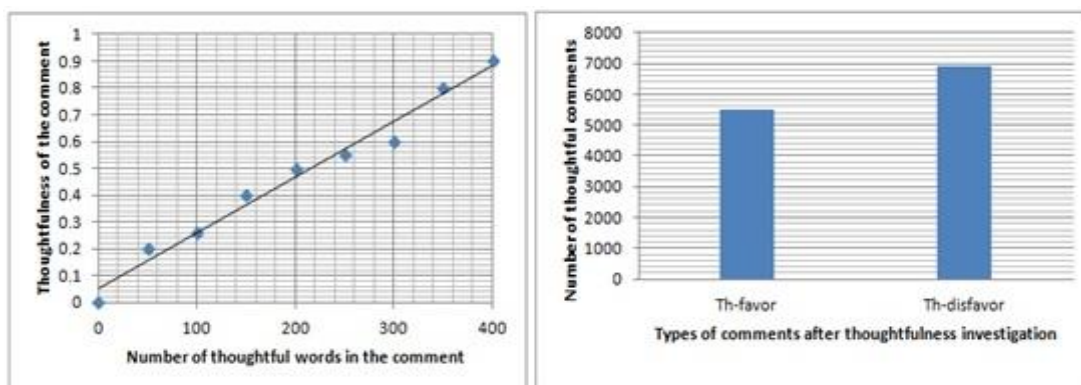


Figure 8. (a) Number of words in the comment and thoughtfulness of the comment (b) Types of comments after thoughtfulness investigation versus number of comments

2.1. Brand popularity predictor

The brand popularity predictor forecasts the Popularity Index PI of the brand based on the brand metadata i.e., N_l, N_s, N_{TFC} (Number of TFCs), and N_{TDFC} (Number of TDFCs). Weights are assigned to each of the metadata empirically, for e.g., $\langle N_l, Low \rangle$, $\langle N_s, Medium \rangle$, $\langle N_{TFC}, High \rangle$, and $\langle N_{TDFC}, High \rangle$. The brand owners initiates several promotional events related to new releases like adding BM videos, creating recent releases cover page, frequently updating status of brand, using hash tags with the product names, timely response to customers, asking questions related to various models of the brand, conducting contests on the new model launch, creating a polling session over the brand posters, and so on. The brand attributes $ai \in \{N_l, N_s, N_{TFC}, \text{and } N_{TDFC}\}$ influences the popularity of the BM on occurrence of Event j at time t , $E_{j,t}$, the Influence Value $IV(a_i, BM, E_{j,t})$ is given by $wa_i * \sum_{k=1}^{k=l} (val(a_i, BP_k, E_{j,t}))$. The Positive Popularity Index of the BM $PPI(BM, E_{j,t}) = \sum_{i=1}^{i=3} [IV(a_i, BM, E_{j,t}) + PPI(RBM, E_{j,t})]$ is influenced by N_l, N_s, N_{TFC} and PPI of the Related Brand Models (RBM), where the $PPI(RBM, E_{j,t})$ is given by $\sum_{K=1}^{K=n} \sum_{i=1}^{i=3} \left[\frac{IV(a_i, RBM_k, E_{j,t})}{3.0} \right] / N_{RBM}$. The Negative Popularity Index of the BM $NPI(BM, E_{j,t}) = IV(a_4, BP, E_{j,t}) + NPI(RBM, E_{j,t})$ is influenced by N_{TDFC} , and NPI of RBM s, where the $NPI(RBM, E_{j,t})$ is given by $\sum_{k=1}^{k=n} IV(a_4, BP, E_{j,t}) / N_{RBM}$. As the Brand B contains several BM s, the PPI and NPI of each of the BM upon $E_{j,t}$ is considered to determine the overall PPI and NPI of the brand, therefore the $PPI(BM, E_{j,t}) = 1/n \sum_{i=1}^{i=n} PPI(BM, E_{j,t})$ and $NPI(BM, E_{j,t}) = 1/n \sum_{i=1}^{i=n} NPI(BM, E_{j,t})$. The overall PI of the B upon $E_{j,t}$ is determined by comparing the PPI and NPI of B upon $E_{j,t}$.

$$PI(B, E_{j,t}) = \begin{cases} High, & \text{if } PPI(B, E_{j,t}) > NPI(B, E_{j,t}) \\ Neutral, & \text{if } PPI(B, E_{j,t}) = NPI(B, E_{j,t}) \\ Low, & \text{if } PPI(B, E_{j,t}) < NPI(B, E_{j,t}) \end{cases}$$

A hybrid time series model with ARIMA (Autoregressive Integrated Moving Average) and neural network [30], [31] is used to compute Forecast Popularity Index of Event j at time $t + k$, $FPI(B, E_{j,t+k})$. The ARIMA prediction model is used to determine the linear relationship over the past observations of popularity index and the neural network model is used on the residues of the ARIMAs output to predict the non-linear relationship in the past observations of PI , a $100*50*10$ neural network model used for popularity prediction is shown in Figure 9. In the first stage, the ARIMA popularity index is computed $API(B, E_{j,t+k}) = \sum_{i=1}^{i=p} [\alpha_i * PI(B, E_{j,t-i})] + \sum_{i=1}^{i=q} [\beta_i * PI(B, E_{j,t-i})] + \epsilon(t)$, where α_i and β_i are empirical constants. Then by using computed $API(B, E_{j,t+k})$, the residual output of ARIMA is calculated $R(B, E_{j,t}) = PI(B, E_{j,t}) - API(B, E_{j,t})$. In the second stage, the neural network popularity index is calculated using n $R(B, E_{j,t})$ outputs $NNPI(B, E_{j,t+k}) = f(R(B, E_{j,t}), R(B, E_{j,t-1}), \dots, R(B, E_{j,t-n})) + \epsilon_i$, where f is a non-linear neural network function, ϵ_i is the random error and the final computed forecasted popularity index is obtained $FPI(B, E_{j,t+k}) = API(B, E_{j,t+k}) + NNPI(B, E_{j,t+k})$.

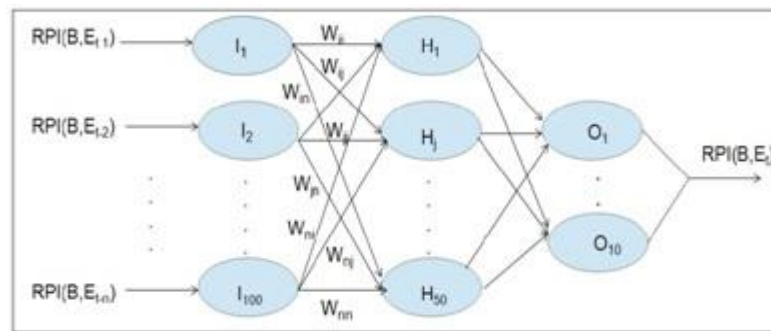


Figure 9. 100*50*10 neural network model for prediction

2.2. Execution time analysis of the proposed framework

The total execution time T_{total} is computed using preprocessing time T_p , relevance checking time T_{rc} , and thoughtfulness investigation time T_{ti} i.e., $T_{total} = T_p + T_{rc} + T_{ti}$. Since, the proposed

framework make use of MapReduce parallel programming model, the T_p is given by the sum of mapping time T_m , shuffling time T_s , and reducing time T_r i.e., $T_p = T_m + T_s + T_r$. The $T_m = \sum_{m=1}^M \left[\frac{CL_m * CS_m}{CP_m} \right]$, where, CL_m is the comment load on map operation, CS_m is cost of map operation, and CP_m is the capacity of map operation. The $T_s = \sum_{i=1, j=1}^{M, S} [CL_m^{op} * M] / CP_s$, where CL_m^{op} is the comment load output of map operation, M is the number of mapping units, CP_s is the capacity of shuffle operation, and S is the number of shuffling units. The $T_r = \sum_{j=1, k=1}^{S, R} [CL_s^{op} * S] / CP_r$, where CL_s^{op} is the comment load output of shuffle operation, CP_r is capacity of reduce operation, and R is the number of reducing units. The time to compute lexical density of comment T_{ldc} , cost of relevance checker CS_{rc} , and capacity of relevance checker CP_{rc} is used to compute $T_{rc} = \sum_{q=1}^C [T_{ldc}^q] * CS_{rc} * CP_{rc}$, where C is the number of preprocessed comments. The $T_{ti} = \sum_{i=1}^T [T_{tr}^i] + T_{ts}$, where T_{tr} is the training time, T_{ts} is the testing time, and T indicates the number of training samples.

2.3. Parallel efficiency of the proposed framework

The parallel efficiency of MapReduce in the proposed framework $E = CT / (M * CT_p)$, where, CT is the total computation time, and CT_p is the computation time on P machines. Consider a single machine environment, where a general computation task is performed on CL of size S . The $CT = (T_m S) + (T_s CL_m^{op} S) + (T_r CL_s^{op} S)$, which is decomposed into map, shuffle, and reduce stages. At first $T_m S$ computation is performed in map stage to yield $CL_m^{op} S$ as output. Then $T_s CL_m^{op} S$ computation is performed in shuffle stage to yield $T_s CL_s^{op} S$ as output. Lastly $T_r CL_s^{op} S$ computation is performed in reduce stage to yield $T_s CL_r^{op} S$ output.

Consider map, shuffle, and reduce stages on P machines parallel environment, where every machine performs map, shuffle, and reduce operations. The $CT_p = \frac{T_m S}{P} + \frac{T_s CL_m^{op} S}{P} + \frac{T_r CL_s^{op} S}{P} + \frac{T_s CL_r^{op} S}{P}$, in which the map stage produces $\frac{CL_m^{op} S}{P}$ as output, shuffle stage produces $\frac{CL_s^{op} S}{P}$ as output, and reduce stage produces $\frac{CL_r^{op} S}{P}$ as output.

Example: Let there be N *BM*s, X comments on every *BM*, and Y words in every comment. The MapReduce model performs Comment preprocessing (C_{pr}), gram count based Shuffling (S_{gc}), and Duplicate Comments removal (DC_r) operations. The calculated $CT = NX Y C_{pr} S_{gc} DC_r$, and $CT_p = \frac{C_{pr}}{P} + \gamma \frac{S_{gc}}{P} + \delta \frac{DC_r}{P}$, where γ is the time taken to read partial output from every mapper and δ is the time taken for shuffling. The parallel implementation of MapReduce is scalable because the efficiency increases with the increase in comments log size and the number of machines.

3. RESULTS AND ANALYSIS

This section provides information about the experimental results of the proposed work in three stages, first the brand metadata source and duration of data collection is indicated, second the experimental setup of OpenStack environment for experiment purpose is discussed and third the performance of the proposed work is evaluated in two stages one is with respect to thoughtful comments identification and other is with respect to brand popularity prediction.

3.1. Hyundai brand metadata

The comment's on Hyundai brand is obtained from the Edmund website which is publicly available at [11]. It is one of the popular American online resource repositories for automobile information, which provides information about car events, dealers, reviews, ownership and so on. The Hyundai car data set consisting of around 50 categories with several heterogeneous features during the span 1/1/2016 to 1/3/2017 is considered for evaluation purpose. More precisely three varieties of comments were extracted like uni-gram, bi-gram, and multi-gram. In order to limit the boundlessness and higher order spanning nature of the user comments, the comments undergo power transformation process. The symbolic features of the comments are normalized using Naive Bayes encoding technique which efficiently handles the multiple categories of comments. In addition, LDA method is used to extract brand relevant words, and even get the rate of relevant and irrelevant words [32].

3.2. Experimental setup

The proposed work is validated using OpenStack private cloud test bed on Open Cirrus test bed available in HP lab website [33], [34]. In order to process the large volume of comment's the Solid State Drives (SSD) is combined with the Hard Disk Drives (HDD) to accelerate the comment's preprocessing and

classification rate. The comments are processed in parallel from HDFS (Hadoop Distributed File System) and the classified comments are written back to HDFS. The performance is evaluated using Apache Hadoop-3.1.0 composed of HDD and SSD, which is capable of performing read, shuffle and reduce operations on 1MB of brand comments. To handle the variability in the metadata, four-fold validation is carried out by doing different levels of partitions and then rounding them off to estimate the popularity of the brand. The experiment is repeated for 10 times using different seed values in order to achieve the accurate prediction value. The popularity prediction is considered as a binary classification problem, the main goal is to predict whether the popularity of the brand is low or high in the near future. The contribution of every trait of the prediction model is estimated using Xboost package in R, and the ARIMA prediction model is trained with learning rate=0.00152, sampling size=0.6, size of the network=100*50*10, and the number of iterations is chosen by four-fold validation technique. The identification of thoughtful favor or disfavor comments added more power to the prediction accuracy. In literature, separate works were found with respect to thoughtful comments identification and popularity prediction. But in the proposed work we do both thoughtful comments identification and popularity prediction, so the performance of the proposed prediction model is compared with the existing works [12], [17] in two perspectives, one is towards thoughtful comments identification and other is towards the brand popularity prediction.

3.3. Thoughtful comments identification

Here the performance of the proposed thoughtful comments identification framework is compared with the existing work discussed in [12]. The comparison is made with respect to parameters like thoughtful comment's identification accuracy, catch rate and miss rate of thoughtful comment's, and execution time. A graph of time versus thoughtful comment's identification accuracy is shown in Figure 10. It is observed from the graph that the accuracy in identifying the thoughtfulness of the comments is higher in the proposed work compared to the existing one. The proposed work is trained to identify thoughtful comment's based on the occurrence of four thoughtful bags of words E, R, C, and Q; most of the target specific commenting terms in social media will be either in positive or negative inclination of the considered bag of words as a result, the thoughtful comment's identification accuracy increases over time. But the existing work relies on logistic model for comment's classification, as the model deals with a lot of independent sentiment words, they are vulnerable to overfitting problem, therefore, the thoughtful comment's identification accuracy decreases.

A graph of comment's log size versus thoughtful comment's catch rate, miss rate of proposed work and existing work are shown in Figure 11 (a) and Figure 11(b). It is observed from the graph that the performance of the proposed work with respect to the catch rate and miss rate of thoughtful comments is found to be good. The accuracy in identifying the thoughtfulness comment's in the proposed work is high as only high relevance comment's are considered for thoughtfulness investigation and lengthier comment's with very low relevance factor are discarded from evaluation after topic modeling, this feature exhibits high correlation with the quality of the comment which influences the catch rate of the thoughtful comment to be higher and the miss rate to be lower. However, in the existing work, the thoughtful comment's identification accuracy is lower as it uses KL divergence algorithm for comment's classification, which exhibits very low correlation with the quality of the comments. Hence, the miss rate of the thoughtful comment is higher and catch rate is lower.

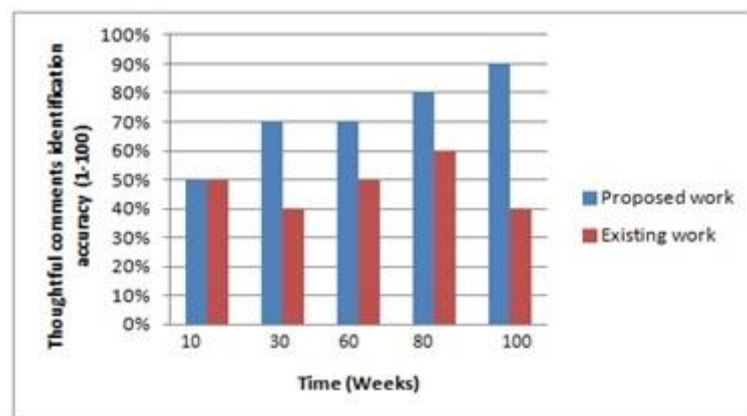


Figure 10. Time versus thoughtful comments identification accuracy

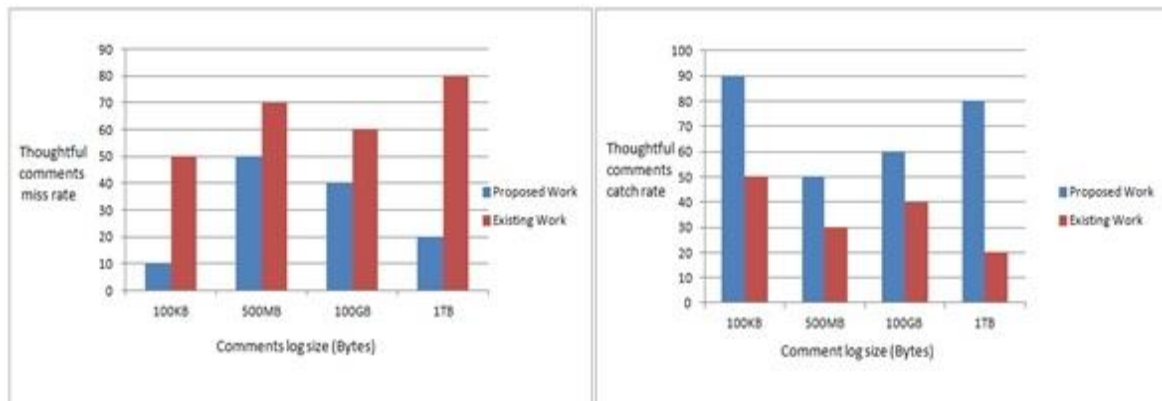


Figure 11. (a) Comments log size versus thoughtful comments catch rate (b) Comments log size versus thoughtful comments miss rate

A graph of comment's log size versus execution time is shown in Figure 12. It is observed from the graph that the total execution time of the proposed thoughtful comments identifier framework is lower compared to the existing work. The proposed framework uses parallel programming model (MapReduce) for comment's preprocessing and thoughtful words sequence matching approach for thoughtful comments identification, which speeds up the execution of the framework with the increase in comment's log size. Whereas in the existing work sequential steps are followed for comment's cleaning, topic modeling, discourse relationship estimation, and logistic regression which increase the execution time with the increase in comment's log size.

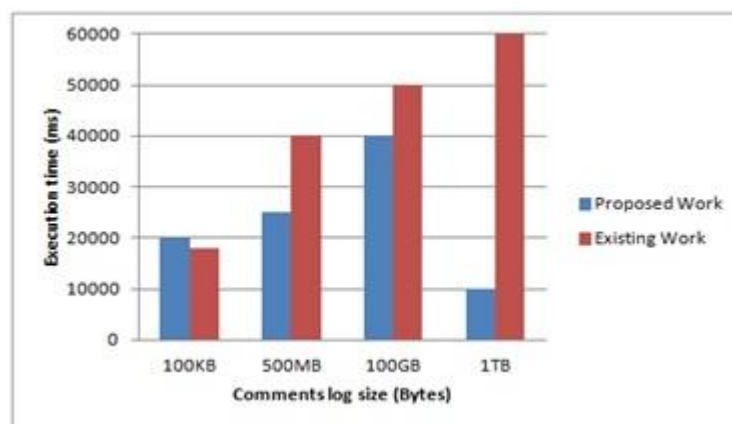


Figure 12. Comments log size versus execution time

3.3. Popularity prediction

Here the performance of the proposed popularity prediction model is compared with the existing work discussed in [17]. The comparison is made with respect to parameters like prediction efficiency, prediction accuracy, and prediction time. A graph of training epochs versus efficiency of the various prediction models is shown in Figure 13. The proposed work uses ARIMA-Neural-Network (ARIMA-NN) model for prediction purpose, the efficiency of it is compared with the well-known prediction model like Neural Network (NN), Gradient Boosting Tree (GBT), and ARIMA. It is observed from the graph that the efficiency of the NN model is very low as it identifies only non-linear relationships in the brand metadata, the efficiency of the GBM is moderate as it cannot extrapolate to unknown relationships of metadata samples, the efficiency of the ARIMA also falls in moderate range as it identifies only linear relationship in the brand metadata, whereas the efficiency of ARIMA-NN is higher because it is able to identify linear, non-linear, and unknown relationships in the brand metadata.

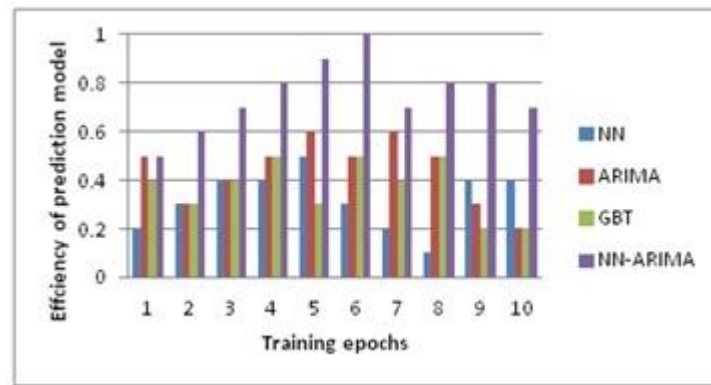


Figure 13. Training epochs versus Efficiency of the prediction models

A graph of time versus prediction accuracy is shown in Figure 14. The prediction accuracy of the proposed work is higher than the prediction accuracy of the existing work. In the proposed work, the comments are not taken as it is instead they are preprocessed, checked for relevance, and then evaluated with respect to thoughtfulness metric, this increases the popularity prediction accuracy with the increase in the number of metadata samples. But in the existing works, the comment's are not filtered i.e., even not-thoughtful comment's are considered while predicting the popularity hence the accuracy of prediction drops with the increase in the number of metadata samples.

A graph of metadata samples versus prediction time is shown in Figure 15. The prediction time of the proposed work is lower as the ARIMA-NN model can be easily generalized to unknown metadata samples during the training phase. But the prediction time of the existing work is higher as the gradient boosting tree model requires careful re-tuning of input parameters whenever unknown relationships appear among the metadata samples.

The performance of the proposed work is found to be good in two aspects one is towards thoughtful comments identification and other is towards popularity prediction. With respect to thoughtful comments identification, the performance is found to be good in terms of parameters like thoughtful comments identification accuracy, catch rate and miss rate of thoughtful comments, and execution time. With respect to popularity prediction, the performance is found to be good in terms of prediction accuracy, prediction efficiency, and prediction time.

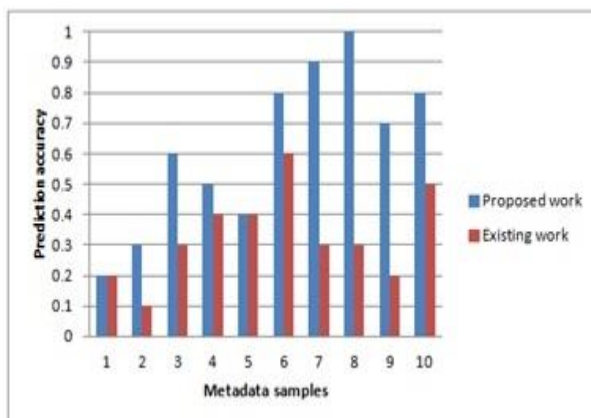


Figure 14. Time versus prediction accuracy

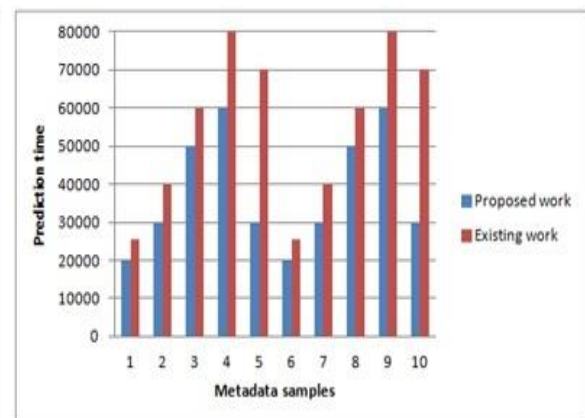


Figure 15. Metadata samples versus prediction time

4. CONCLUSION

This paper forecast the brand popularity by analyzing the metadata of the brand. The working of proposed brand predictor model is tested on Hyundai cars comment log of 120 weeks. Compared to the recent existing works, the performance of the proposed work is found to be encouraging with respect to

thoughtful comment's identification accuracy, thoughtful comment's identification rate, prediction time, prediction accuracy, and efficiency of the prediction model.

REFERENCES

- [1] S. Assegaff, *et al.*, "Social Media Success for Knowledge Sharing: Instrument Content Validation", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, pp. 2447-2453, 2016.
- [2] L. Phillips, *et al.*, "Using Social Media To Predict the Future: A Systematic Literature Review", arXiv:1706.06134, 2017.
- [3] H. Bagheri and A. A. Shaltooli, "Big Data: Challenges, Opportunities and Cloud Based Solutions", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, pp. 340-343, 2015.
- [4] "Social media", <http://government-2020.dupress.com/driver/social-media/>, Accessed, 12-Oct-2017.
- [5] "Marketing Media Money", <https://www.cnbc.com/2016/12/05/social-media-advertising-spend-set-to-overtakenewspapers-by-2020-research.html>, Accessed, 12-Oct-2017.
- [6] <https://www.reuters.com/article/us-advertising-forecast-idUSKBN13U001>, Accessed, 12 Oct 2017.
- [7] S.T.M. Kumar and S.K. Singh, "A Review of Social Media: In Future", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 8, December 2017.
- [8] K. Curran, *et al.*, "Google+ v Facebook: The Comparison", *Telecommunication, Computing, Electronics and Control (TELKOMNIKA)*, vol. 10, 2012.
- [9] T. Vukasovic, "Brand Developing Relationships through Social Media", *International Conference on Management Knowledge and Learning*, vol. 21, pp. 97-105, 2013.
- [10] A. Davoudi, "Customer Misuse of Social Media and Consequences on Firm Strategies", *5th IBA Bachelor Thesis conference*, 2015.
- [11] "Edmunds", <https://www.edmunds.com/>, Retrieved, 12-Oct-2017.
- [12] S. Gottipati, S and J. Jiang, "Finding Thoughtful Comments from Social Media", *24th International Conference on Computational Linguistics*, vol. 12, pp. 995-1010, 2012.
- [13] J. G. Lee, *et al.*, "An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors", *International Conference on Intelligence and Intelligent Agent Technology*, vol. 1, pp. 623-630, 2010.
- [14] H. Lakkaraju and J. Ajmera, "Attention Prediction on Social Media brand pages", *20th ACM international Conference on Information and Knowledge Management*, pp. 2157-2160, 2011.
- [15] R. Bandari, *et al.*, "The Pulse of News in Social Media: Forecasting Popularity", *Sixth International AAAI Conference on Weblogs and Social Media*, pp. 26-33, 2012.
- [16] S. Gottipati, "Opinion Mining of Sociopolitical Comments from Social Media", Institutional Knowledge at Singapore Management University, 2014.
- [17] T. Uddin, *et al.*, "Predicting the Popularity of Online News from Content Metadata", *International Conference on Innovations in Science, Engineering and Technology*, 2016.
- [18] K. Fernandes, *et al.*, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", *Portuguese Conference on Artificial Intelligence*, pp. 535-546, 2015.
- [19] Z. Liu, "Statistical Models to Predict Popularity of News Articles on Social Networks", Arts and Sciences Electronic Theses and Dissertations, 2017.
- [20] Y. Keneshloo, *et al.*, "Predicting the Popularity of News Articles", *International Conference on Innovations in Science, Engineering and Technology*, 2017.
- [21] M. Zudrell, "Factors Affecting branded Posts Popularity and fan page Engagement", Thesis submitted to Modulvienna University-WKO-Private University, 2016.
- [22] J. Heijden, "Facebook forwards", Thesis submitted to Tilburg university, 2013.
- [23] S.A. Thanekar, *et al.*, "Big Data and MapReduce Challenges, Opportunities and Trends", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, 2016.
- [24] A. Shrivastava and B. Pant, "Opinion Extraction and Classification of real time Facebook Status", *Global Journal of Computer Science and Technology*, vol. 12, 2012.
- [25] C. Virmani, *et al.*, "Extracting Information from Social Network using NLP", *International Journal of Computational Intelligence Research*, vol. 13, 2017.
- [26] P. K. Kumar and S. Nandagopalan, "Insights to Problems, Research Trend and Progress in Techniques of Sentiment Analysis", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, 2017.
- [27] M. Injadat, *et al.*, "Data Mining Techniques in Social Media", *Journal of Neurocomputing*, vol. 214, pp. 654-670, 2016.
- [28] M. Steedman, "Some Important Problems in Natural Language Processing", Informatics Hamming Seminar, 2010.
- [29] B. H. Nesia and S. A. Ginting, "Lexical Density of English Reading Texts for Senior High School", *Transform Journal of English Language Teaching and Learning of FBS UNIMED*, vol. 3, 2014.
- [30] A. Verma, *et al.*, "Analysis of Time-series Method for Demand Forecasting", *Journal of Engineering and Applied Sciences*, vol. 12, pp. 3102-3107, 2017.
- [31] S. Celik, *et al.*, "Forecasting the Production of Groudnut in Turkey using ARIMA Model", *The Journal of Animal and Plant Sciences*, vol. 27, pp. 920-928, 2017.
- [32] H. Jelodar, *et al.*, "Latent Dirichlet allocation (LDA) and Topic Modeling: Models, Applications, a Survey", arXiv:1505.07302, 2015

- [33] A. I. Avetisyan, *et al.*, "Open Cirrus: A Global Cloud Computing Testbed", *IEEE Computer Society*, pp. 35-43, 2010.
- [34] "mloss.org", <http://mloss.org/software/view/543/>, Accessed, 12-Oct- 2017.

BIOGRAPHIES OF AUTHORS



Bhargavi K received her bachelors and masters degree in computer science and engineering from visveswaraya Technological University (VTU). She is currently pursuing Ph.D. under VTU; her research interest includes application of cognitive agents in healthcare, developing context aware smart applications, converting SQL queries to XML, high performance computing, swarm intelligence, and machine learning.



B.Sathish Babu received his Ph.D degree in Electrical Communication Engineering from Indian Institute of Science, Bangalore, India. His research interest includes Cognitive agents based control solutions for Networks, Grid Computing, Cloud Computing Scheduling and Security Issues, Context-aware Trust issues in Ubiquitous Computing, High performance computing, Privacy issues in WSN, and Opportunistic Computing. His publications includes some of the reputed journals like Elsevier, Wiley, CRC Press, InderScience, IGI, PHI, TMH, etc.



S. S. Iyengar is a Distinguished Ryder Professor and Director of the School of Computing and Information Sciences at Florida International University, Miami. Dr. Iyengar is a pioneer in the field of distributed sensor networks/sensor fusion, computational aspects of robotics and high performance computing. He has published over 600 research papers and has authored/edited 22 books published by MIT Press, John Wiley & Sons, Prentice Hall, CRC Press, Springer Verlag, etc.